DOCUMENT RESUME

ED 194 564 TM 800 693

AUTHOR Jaeger, Richard M.
TITLE Some Exploratory Indices for Selection of a Test Equating Method.
PUB DATE Apr 80
NOTE 40p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations: Difficulty Level: *Equated Scores: Higher Education: *Item Analysis: *Scaling: *Scoring Formulas: Test Construction: Testing Problems: Test Items
IDENTIFIERS *Equipercentile Equating: *Linear Equating Method: Test Equivalence

ABSTRACT

Five statistical indices are developed and described which may be used for determining (1) when linear equating of two approximately parallel tests is adequate, and (2) when a more complex method such as equipercentile equating must be used. The indices were based on: (1) similarity of cumulative score distributions: (2) shape of the raw-score to scaled-score transformation: (3) consistency of linear and equipercentile equating results: (4) similarity of item difficulty distributions: and (5) similarity of item discrimination distributions. Data used in the study were collected from between 5000 and 6000 examinees participating in a national aptitude testing program used for college admission selection purposes. Eight forms of the test were used in five separate equatings over three years. Four of five indices of the adequacy of linear equating distinguish markedly between two equatings for which the linear equating method was demonstrably inadequate, and three equatings for which it appeared to be satisfactory. The index based on similarity of item difficulty distributions appeared to be the most useful and discriminated very clearly between problematic and satisfactory linear equatings. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194564

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Some Exploratory Indices for Selection of a Test Equating Method

Richard M. Jaeger

University of North Carolina at Greensboro

"PERMISSION TO REPHODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Jaeger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800693

Running head: Some Exploratory Indices for Selection

Presented at the 1980 Annual Meeting of the American Educational Research Association, April 7-11, Boston, Mass.

Abstract

Five statistical indices for determining when linear equating of two approximately parallel tests is adequate, and when a more-complex method (such as equipercentile equating) must be used, are developed and described. When used with data from five test equatings from a national aptitude testing program, four of the five indices were found to discriminate between two problematic equatings, and three for which linear equating was found to be adequate using external evidence. Implications for practice in large-scale testing programs are discussed.

Some Exploratory Indices for Selection of a Test Equating Method

To preserve test security, to sustain the integrity of claims to generalizability, and for a host of other, equally sound reasons, most major testing programs require the creation of multiple test forms. Because multiple test forms, regardless of the care with which they are constructed, invariably differ in difficulty level and range, some method must be used to translate scores on one form of the test to the scale of another. When the test forms are essentially parallel, such translations are termed "test equating."

The methodology of test equating has enjoyed an abundant literature since the early 1950's (Lord, 1950; Flanagan, 1951; Lord, 1955), with periodic resurgence of interest about twice per decade since (Angoff, 1961; Lennon, 1964; Angoff, 1968; Maxey and Lenning, 1974). Angoff (1971) provides an excellent summary of the methodological literature to that time. More-recent articles have concentrated on the application of latent trait models to problems that include test equating through so-called "sample-free" specification of test item parameters.

Neglecting the latter work for reasons of practicality which will not be discussed here, two forms of test equating have been recommended or assumed throughout the three decades of work cited -- either equipercentile equating or linear equating. The equipercentile method derives from a basic definition of equivalent scores that was provided by Lord (1950) and Flanagan (1951), and is cited by Angoff (1971, p. 563):

"A commonly accepted definition of equivalent scores is: Two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal."

The linear equating method is based on the assumption that two forms of a test that are constructed so as to be parallel -- that is, designed, to the best abilities of the test constructors, to "measure exactly the same thing in the same scale and, in a sense, measure it equally well for all persons." (Lord and Novick, 1968, p. 48) — will have raw-score distributions that are virtually identical, apart from possible differences in mean and variance. Given this assumption, it should be possible to translate scores on one form to the scale of another form simply by equating standard scores. In other words, if the raw-score distributions of forms X and Y differ only in their first two central moments, scores on Form Y can be placed on the scale of Form X by using the linear transformation:

$$\frac{Y - M_y}{s_y} = \frac{X - M_x}{s_x}, \qquad (1)$$

where M_x and M_y are respective means of the X and Y raw-score distributions, and s_x and s_y are respective standard deviations of these distributions.

In discussing the linear equating method, Roman letters have been used to symbolize the means and standard deviations because estimates of the corresponding population parameters, rather than the parameters themselves, will be available in any practical application.

The problem treated in this paper, when to use the linear equating method and when to use the equipercentile method, has been virtually ignored in previous methodological literature. Although previous articles have clearly delineated the theory underlying both methods and the differences between them, choice of method has been left largely to assumption or conjecture. For example, Lord (1955; p. 193) states:

> "Two equally reliable tests may be equated by changing the score scale on either test in such a way that the distribution of actual scores becomes the same for both tests. The equipercentile method of equating is commonly used for this purpose. If tests are equally reliable and approximately parallel, it is often assumed that score distributions differ in mean and variance, but any other differences can be ignored."

Lord then goes on to provide efficient estimators of the mean and variance parameters needed to effect the linear transformation specified in Equation (1). Note that he does not treat the problem of how "approximate" parallelism need be, to support the assumption of sufficiency of the linear equating method.

Angoff (1971; p. 564) makes an observation similar to Lord's:

> "It is important to emphasize that linear equating is a very close approximation to equipercentile equating when the shapes of the raw score distributions are similar. If one is prepared to assume that differences in the shapes of the distributions of raw scores on the two forms are sufficiently trivial so they may be disregarded, linear equating is to be preferred. Unlike equipercentile equating, it is entirely analytical and verifiable and is free from any

errors of smoothing, which can produce serious errors in the score range in which data are scant and/or erratic."

Angoff's latter point gets to the heart of the practical problem in choosing between the linear and equipercentile equating methods. When test forms are not perfectly parallel, the two methods will produce somewhat different results. These differences are likely to be most noticeable at the extremes of the score distribution, regions with relatively sparse data, but still of critical interest in many testing applications. For example, when tests are used to aid in selection decisions, the most selective institutions will be interested in accurate discrimination among applicants with extremely high scores. Interest in the extreme low end of the score scale arises in the minimum competency testing programs that are currently flourishing in many states and local school systems (Jaeger and Tittle, 1980).

In this paper, we discuss five indices that should logically discriminate between situations where the linear equating method adequately adjusts for differences between the score distributions of two approximately parallel test forms, and situations where a more complex model is needed. The indices are applied to data resulting from the administration of eight different forms of a widely-used aptitude test in an ongoing national testing program. The data were collected over a span of three years, in five different test equating studies.

Method

Data Source

The data used in this study were collected during the standard

administrations of an aptitude test that is used for admission selection purposes by many colleges and universities throughout the nation. As a part of the ongoing testing program, new forms of the test are routinely equated to previously-used forms, in order to maintain form-to-form comparability both within and across years. Eight forms of the test, here designated T1 through T8, were used in five separate equatings over a three-year time period. The procedure that Angoff (1971, p. 569) terms "Design I: Random Groups" was used to collect equating data. In that design, independent random samples of examinees complete each of the forms to be equated. No attempt is made to adjust for between-sample differences through use of common test items or an anchor test. Sample sizes varied somewhat across equatings, but somewhere between 5000 and 6000 examinees completed each form of the test.

In two of the test equatings used for this study, preliminary evidence suggested that the linear equating results might be faulty. When the linear method was used and equated raw scores were transformed to a standard score scale, about eight to ten times the number of examinees typically found, scored at the extreme upper limit of the score scale. In the other three equatings, numbers of extreme high-scorers approximated those typically found over a period of many years in nearly 100 test equatings. These results suggested that, for the purposes of the testing program, the linear equating method was successful in maintaining the integrity of the standardized score scale in the latter three equatings.

Indices and Exploratory Analyses

Similarity of Cumulative Score Distributions. As an index of the

adequacy of the linear equating procedure, cumulative raw-score distribution functions of "new" and "old" equated tests were compared, following adjustment for differences in means and standard deviations. The Kolmogorov-Smirnov two-sample test of the equality of cumulative distribution functions was used to effect the comparison (Smirnov, 1948).

The following procedure was applied to the pairs of tests involved in each of the five equatings. First, linear equating parameters were estimated and applied to translate raw scores on the new test to the scale of the old test. Lord's (1955) maximum likelihood solution was used for parameter estimation. Following transformation, mid-percentile ranks were computed for equally-spaced scores on the new test. Third, linear interpolation was applied to the cumulative distribution function of raw scores on the old test, to determine mid-percentile ranks on the old test, corresponding to translated raw-score values on the new test. Finally, differences in mid-percentile ranks were computed for use in the Kolmogorov-Smirnov two-sample test. Differences in mid-percentile ranks were computed for scores on the new test in multiples of five, for which mid-percentile ranks could be computed on both the old test and the new test. Because of the nature of the linear equating method, the ranges of raw scores on the old test and translated raw scores on the new test typically do not overlap completely. Thus corresponding mid-percentile ranks could not be computed over the complete ranges of both tests. In the equating of Test T7 to Test T6, for example, corresponding mid-percentile ranks were computed for scores of 25, 30, 35, . . ., 140, on the transformed raw-score scale of Test T6.

The Kolmogorov-Smirnov two-sample test provides a critical value for the maximum difference between corresponding mid-percentile ranks of two variables. If the difference between the cumulative distribution functions of raw scores on the old test and linearly equated raw scores on the new test differ by an amount that exceeds the critical value of the Kolmogorov-Smirnov test for some reasonably specified Type I error level, it might be argued that a more complex equating method is required.

Shape of the Raw-Score to Scaled-Score Transformation. If the linear equating procedure adequately accounts for differences in the raw-score distributions of two test forms, it is reasonable to assume that an equipercentile transformation of the raw scores of the new form to the scale of the old form will be essentially linear. This logic was employed in developing a second index of the adequacy of linear equating.

First, the equipercentile equating method was used to translate raw scores on the new form of each test pair to the standard-score scale of the old form in each pair. Equipercentile scaled scores on the new forms were then used as dependent variables in third-order stepwise polynomial regressions on the raw scores of the new forms. For example, in the equating of Test T3 to Test T1, the equipercentile equating method was used to compute scores on Test T3 on the standard-score scale of Test T1. These equated scaled scores on Test T3 were then regressed successively on 1) raw scores on Test T3, 2) raw scores and squared raw scores on Test T3, and 3) raw scores, squared raw scores and cubed raw scores on Test T3. Similar procedures were applied to the data in each of the other four equatings.

Coefficients of determination and increments in coefficients of determination were computed at each step in the stepwise regression analysis. In addition, the increment in the coefficient of determination between the first and third steps of each regression analysis was computed.

The increment in the coefficient of determination between Step 1 (with raw scores alone as the independent variable) and Step 3 of each regression analysis (with raw scores, squared raw scores and cubed raw scores as independent variables) is proposed as an index of the adequacy of linear equating. If the linear method adequately equates the score distributions of two test forms, the linear and equipercentile methods should produce very similar results. Since the linear equating method can be defined solely in terms of a linear regression of scaled scores on raw scores, similarity between the results of the linear and equipercentile methods can be expressed in terms of the degree to which equipercentile scaled scores can be similarly defined. Thus the linear equating method can be described as adequate if a regression of equipercentile scaled scores on raw scores has a coefficient of determination that is essentially one and if the addition of squared raw scores and cubed raw scores to the regression equation provides an increment in the coefficient of determination that is essentially zero.

Unfortunately, formal statistical hypothesis testing is of little value in determining how large an increment in the coefficient of determination can still be viewed as "essentially zero." Because the sample sizes involved in adequately precise test equatings are very large, a

miniscule increment in the coefficient of determination would probably be statistically significant at any reasonable Type I error level.

Consistency of Linear and Equipercentile Equating Results. A third approach to determining the adequacy of linear equating involved a direct comparison of the results of linear equating and equipercentile equating. The analysis was based on the assumption that an equipercentile equating function would fluctuate randomly around a corresponding linear equating function, if the linear equating method was adequate.

To explore the deviation of equipercentile equating results from corresponding linear equating results, both methods were applied to five pairs of tests. For each raw score on the new test of a pair, the scaled score determined through the equipercentile method of equating was subtracted from the corresponding scaled score determined through the linear equating method. These difference scores were then used as dependent variables in a set of polynomial regression analyses on raw scores, just as in the analyses described immediately above. It was reasoned that random deviations of equipercentile scaled scores from corresponding linear scaled scores should not be predictable from any function of raw scores, and that a third-order polynomial in raw scores approximated the first few terms of a Taylor series in the functional relationship between the deviations and raw scores on the new test form.

Once again, coefficients of determination were computed at each step of the stepwise regression analyses. In addition, increments in the coefficients of determination were computed between steps 1 and 2, steps 2 and 3, and steps 1 and 3.

Small coefficients of determination at each step would support a claim to the adequacy of the linear equating method, and examination of increments in coefficients of determination would help to describe the functional relationship between linear minus equipercentile scaled scores and raw scores, should such relationships prove to be non-random.

Similarity of Item Difficulty Distributions. A fourth index of the adequacy of linear equating was based on the acknowledged sufficiency of linear equating for test forms that are essentially parallel. The greater the deviation from parallelism, the greater the need for a more-complex equating model, since the score distributions of non-parallel forms are likely to differ in higher-order moments.

One determinant of the score distribution of a test is the distribution of the difficulties of items that compose the test. One would expect parallel test forms to have identical distributions of item difficulties. A potentially useful index of the adequacy of linear equating, then, would be one that reflected the similarity of the cumulative distributions of item difficulties. In keeping with this reasoning, Kolmogorov-Smirnov two-sample tests of the equality of the cumulative distribution functions of transformed item difficulties were conducted for each of the five pairs of test forms. Prior to computing cumulative distribution functions, item p-values were transformed to normalized standard scores (deltas) with a mean of 13 and a standard deviation of four, as described in Angoff and Ford (1973). The test statistic from the Kolmogorov-Smirnov test is proposed as an index of the adequacy of linear equating.

Similarity of Item Discrimination Distributions. A fifth index of the adequacy of linear equating, the test statistic from a two-sample Kolmogorov-Smirnov test of the equality of cumulative distributions of biserial correlation coefficients, is based on the same logic as the fourth index, discussed immediately above. If two test forms are essentially parallel we would expect their distributions of item discrimination indices to be similar if not identical. The lower their degree of parallelism, the greater should be the difference between the distributions of their item discrimination indices.

Two-sample Kolmogorov-Smirnov tests of the equality of distributions of biserial correlation coefficients were conducted for five pairs of tests; resulting test statistics were compared for the two problematic linear equatings and the three satisfactory linear equatings.

Results and Discussion

Similarity of Cumulative Score Distributions. The results of two-sample Kolmogorov-Smirnov tests of the equality of adjusted cumulative raw-score distributions for five pairs of equated tests are summarized in Table 1. Although differences between adjusted raw-score cumulative

---

Insert Table 1 about here

---

distribution functions were not significant at the 0.05 level for any of the five equatings, there is a marked difference between the size of the Kolmogorov-Smirnov test statistic for the two problematic linear equatings and the three satisfactory linear equatings. The maximum differences between raw-score cumulative distribution functions for the two problematic

equatings (Test T7 to Test T6, and Test T8 to Test T6) approach statistical significance, whereas corresponding statistics for the other three linear equatings are far smaller than their respective critical values. Thus the Kolmogorov-Smirnov test statistic appears to distinguish well between pairs of tests that can be equated satisfactorily using the linear method, and pairs of tests for which a more-complex equating method must be used. Unfortunately, these findings do not suggest a statistically-based upper bound for the test statistic, beyond which linear equating should be avoided. Additional empirical work with other tests in other settings will be necessary to develop such a bound.

Shape of the Raw-Score to Scaled-Score Transformation. The relationship between equipercentile scaled scores and raw scores on the test to be equated was explored in two different ways for each of five test equatings. First, to gain insight into the nature of relationships between equipercentile scaled scores and corresponding raw scores, scatter diagrams were drawn with equipercentile scaled scores on the vertical axis and raw scores on the horizontal axis. These are shown in Figures 1 through 5.

Scatter diagrams in Figures 1 through 3, for the three equatings for which the linear model appeared satisfactory (Test T2 to Test T1, Test T4 to Test T1, and Test T5 to Test T2), show an essentially linear relationship between equipercentile scaled scores and corresponding raw scores with the exception of some relatively small deviations at the extremes of the score scale. In contrast, scatter diagrams for the two

---

Insert Figures 1, 2 and 3 about here

---

problematic equatings, Figures 4 and 5, show marked curvilinearity in the relationships between equipercentile scaled scores and corresponding raw scores. These figures suggest the inadequacy of a linear equating model.

Insert Figures 4 and 5 about here

Although they don't provide a specific statistical index, these graphical results lend credence to the idea of examining the relationship between equipercentile scaled scores and corresponding raw scores, when searching for an index of linear equating adequacy.

Results of the second statistical procedure for examining the adequacy of linear equating, described in the Method section above, are shown in Table 2. In this table, attention should be paid to the increments in the coefficient of determination between Step 1 and Step 3 of the stepwise

Insert Table 2 about here

polynomial regressions. In Step 1, equipercentile scaled scores were predicted from raw scores alone, using a simple linear regression equation. In Step 3, equipercentile scaled scores were predicted from a third-order polynomial in raw scores, a function that would accomodate either a parabolic or an S-shaped relationship between equipercentile scaled scores and raw scores. For the two problematic linear equatings (Test T7 to Test T6 and Test T8 to Test T6), increments in the coefficient of determination from Step 1 to Step 3 vary from nearly three times as large to almost fourteen times as large as corresponding increments for the three satisfactory linear equatings.

This statistical index thus appears to discriminate well between test forms for which linear equating is satisfactory and forms for which a more-complex procedure must be used. However, a suitable upper bound on the index is not apparent in these results. Because of the very large sample sizes involved in the equatings, even the smallest increment in the coefficient of determination (.00068 for the equating of Test T3 to Test T1) is significant well beyond the .001 level. Although one could set an upper bound of 0.004 on the basis of these results, such a recommendation would have to be subjected to additional empirical investigation.

Consistency of Linear and Equipercentile Equating Results. Results for the third index of the adequacy of linear equating, based on predictability of the difference between scaled scores derived through linear equating and scaled scores derived through equipercentile equating, are shown in Table 3.

Insert Table 3 about here

As asserted in the discussion above, if linear equating adequately models the relationship between the raw-score distributions of two test forms, the results of linear equating and equipercentile equating should be virtually indistinguishable, apart from random differences. In particular, the difference between scaled scores on the new test form computed using linear equating, and corresponding scaled scores computed using equipercentile equating, should not be systematically predictable from any function of corresponding raw scores.

Data in Table 3 show that this logic was not completely supported by any of the five equatings investigated, regardless of external evidence on the adequacy of linear equating. Using third-order polynomials in raw scores, differences between linear scaled scores and equipercentile scaled scores could be predicted with at least a moderate degree of precision in all five equatings; coefficients of determination for the five equatings varied between 67 percent and 96 percent. It is interesting to note, however, that the level and pattern of predictability is quite different for the two problematic linear equatings (Test T7 to Test T6 and Test T8 to Test T6) and the three satisfactory linear equatings. In the case of the problematic equatings, 96 percent and 94 percent of the variance of the difference variable, respectively, was predicted from a third-order polynomial in raw scores. These figures contrast markedly with predictabilities in the range of 68 percent to 81 percent for the satisfactory linear equatings. In addition, although a linear regression equation afforded minimal predictability in all five equatings, the increment in predictability afforded by a quadratic term in raw scores was considerably higher for the two problematic linear equatings than for the three satisfactory ones. The highest increment for the satisfactory equatings was 71 percent of the variance of the difference variable, while the smallest corresponding increment for the problematic equatings was over 92 percent.

Scatter diagrams of the difference variable and corresponding raw scores for the five equatings, omitted here for the sake of brevity, suggest markedly different patterns of relationship for the problematic

linear equatings and the satisfactory ones. For the problematic cases, relationships between the difference variable and corresponding raw scores are decidedly parabolic over the entire raw-score range. The predictability of the difference variable from a quadratic equation in raw scores is apparent. For the satisfactory cases, the difference variable appears to deviate from zero only slightly (by less than two percent of the range of scaled scores on the test forms) over at least 80 percent of the raw-score scale. Only at the extremes of the raw-score scale does the difference variable deviate systematically and appreciably from zero. Since the pattern of deviation is similar at both ends of the raw-score scale (either consistently positive or consistently negative, depending on the particular test forms being equated), the overall shape of the relationship between the difference variable and the raw-score scale is somewhat parabolic. Hence the moderate degree of predictability from a quadratic function of raw scores, even for the apparently satisfactory linear equatings.

These results suggest that the coefficient of determination resulting from prediction of the difference between linear and equipercentile scaled scores, using a third-order polynomial in raw scores for prediction, might serve as an index of the adequacy of linear equating. Once again, there is no purely statistical basis for defining an upper bound on the coefficient of determination that would clearly prescribe use of a nonlinear equating method. The data in Table 3 would support a cutoff value of 90 percent (That is, do not use linear equating if the coefficient of determination exceeds 0.90.), but such a rule would clearly require

confirmation in a variety of equating situations using a wide range of tests.

Similarity of Item Difficulty Distributions. Data resulting from statistical comparison of the cumulative distribution functions of transformed difficulty values (equated deltas) are shown in Table 4. When maximum differences between cumulative distribution functions for members of the five pairs of tests are compared with corresponding critical values for the two-sample Kolmogorov-Smirnov test (alpha = .001), the equatings for which the linear method was problematic are clearly distinguished from the equatings for which it was satisfactory. Differences in cumulative distribution functions of equated deltas are significant well beyond the 0.001 level for the two problematic equatings. Similar differences for two of the three satisfactory linear equatings are not significant at this probability level, and the maximum difference for the third satisfactory equating barely exceeded the .001 critical value.

Insert Table 4 about here

In addition to providing a purely statistical criterion for distinguishing between apparently satisfactory and problematic linear equatings, this procedure has another distinct advantage. Unlike the indices described above, which require data from a linear equating, an equipercentile equating, or both, this index of the adequacy of linear equating requires only a listing of item difficulty values for each of the tests to be equated. It can thus be applied as a part of a standard test

development procedure, using data from field tryouts of test items. Given this advantage, it should be possible to tailor a new form of a test so that linear equating will provide a satisfactory transformation of the new form to the scale of existing forms. As Angoff (1971; p. 564) has suggested, when it is appropriate, linear equating has the advantage of being a purely analytic procedure. It does not require smoothing of the score distributions of the tests to be equated, nor any other subjective decision-making.

Similarity of Item Discrimination Distributions. The fifth index explored, involving comparison of the cumulative distribution functions of biserial correlation coefficients for the tests to be equated, did not distinguish between the three satisfactory linear equatings and the two that were problematic. As the summary of two-sample Kolmogorov-Smirnov tests (Table 5) indicates, test statistics for the two problematic equatings barely exceeded those of the three satisfactory equatings. And all five test statistics exceeded their corresponding critical values at the .001 level.

Insert Table 5 about here

It would appear that the similarity of distributions of item discrimination indices is not nearly as sensitive an indicant of the parallelism of two test forms as is the similarity of their distributions of item difficulties. Or, alternatively, this aspect of parallelism appears to have little influence on the adequacy of linear equating.

Conclusions

Although the results presented here are preliminary in that they derive from a single testing program and involve only five test equatings, they do provide a promising beginning to the solution of a heretofore unexamined problem. Four of our five indices of the adequacy of linear equating distinguish markedly between two equatings for which the linear equating method was demonstrably inadequate and three equatings for which it appeared to be satisfactory. The fourth index explored, based on the similarity of item difficulty distributions, appeared to be the most useful. It discriminated very clearly between the problematic and satisfactory linear equatings. A purely statistical criterion for rejecting the linear equating method in favor of a more-complex procedure appeared to suffice. And finally, this index can be computed from data gathered in test item tryouts, as part of a standard test development program. This last feature would allow tailoring of new test forms so as to provide satisfactory linear equating.

The problem of choosing a test equating method has plagued those responsible for major testing programs for several decades. The effects of an erroneous decision are serious, as has been demonstrated repeatedly in testing programs of nationwide scale. Yet equating methods are presently chosen on the basis of tradition, habit, or the purely theoretical advantages of one method or another, with little empirical attention to the consequences.

As states and school systems embark on the development of minimum competency tests, their needs for sound test equating procedures will

quickly be realized. Test security demands will certainly lead to the development of multiple test forms, and these will have to be placed on a common score scale if equivalent decisions on passage or failure of individual students are to be made. Thus the problems that motivated this research are certain to become even more important to testing practitioners in the ensuing years.

Further examination of the results presented here -- by applying the first four indices to pairs of equated forms from a variety of testing programs — is both warranted and encouraged. If these results are confirmed in a wide range of settings, a useful tool for test developers will have been identified.

References

Angoff, W. H. Basic equations in scaling and equating. In S. S. Wilks (Ed.), Scaling and equating college board tests. Princeton, N.J.: Educational Testing Service, 1961, 120-129.

Angoff, W. H. How we calibrate college board scores. College Board Review, 1968, 68, 11-14.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement. (2nd ed.) Washington: American Council on Education, 1971, 508-600.

Angoff, W. H. and Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Flanagan, J. C. Units, scores and norms. In E. F. Lindquist (Ed.), Educational Measurement. Washington: American Council on Education, 1951, 695-763.

Jaeger, R. M. and Tittle, C. K. Minimum competency achievement testing: Motives, methods, models and consequences. Berkeley, California: McCutchan (~~In press~~ 1980).

Lennon, R. T. Equating nonparallel tests. Journal of Educational Measurement, 1964, 1, 15-18.

Lord, F. M. Notes on comparable scales for test scores. (ETS RB 50-48) Princeton, N.J.: Educational Testing Service, 1950.

Lord, F. M. Equating test scores -- a maximum likelihood solution. Psychometrika, 1955, 20, 193-200.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Maxey, E. J. & Lenning, O. T. Another look at concordance tables between ACT and SAT. Journal of College Student Personnel, 1974, 15, 300-304.

Table 1

Summary of Results of Two-Sample Kolmogorov-Smirnov Tests of Equality of Cumulative Distribution Functions, Following Linear Equating Transformations, for Five Pairs of Aptitude Tests

Test Equating	Maximum Difference in Cumulative Distributions	Minimum Difference Needed for significance (.05 level)
T3 to T1	0.0079	0.0275
T4 to T1	0.0135	0.0275
T5 to T2	0.0059	0.0248
T7 to T6	0.0204	0.0256
T8 to T6	0.0212	0.0256

Table 2

Summary of Results of Polynomial Regression of Equipercentile Scaled Scores on Linear, Quadratic, and Cubic Raw-Score Terms for the "New-Test" Member of Five Pairs of Aptitude Tests

Test Equating	Coefficients of Determination			Increments in Coefficients of Determination		
	Step 1	Step 2	Step 3	Step 1 to 2	Step 2 to 3	Step 1 to 3
T3 to T1	.99897	.99949	.99965	.00052	.00016	.00068
T4 to T1	.99519	.99831	.99836	.00312	.00005	.00317
T5 to T2	.99918	.99978	.99984	.00060	.00006	.00066
T7 to T6	.99149	.99947	.99964	.00798	.00017	.00815
T8 to T6	.99040	.99935	.99938	.00896	.00003	.00899

Table 3

Summary of Results of Polynomial Regression of the Difference Between Linear and Equipercentile Scaled Scores on Linear, Quadratic, and Cubic Raw-Score Terms for the "New Test" Member of Five Pairs of Aptitude Tests

Test Equating	Coefficients of Determination			Increments in Coefficients of Determination		
	Step 1	Step 2	Step 3	Step 1 to 2	Step 2 to 3	Step 1 to 3
T3 to T1	.04635	.52681	.67595	.48046	.14914	.62960
T4 to T1	.05881	.66903	.67892	.61022	.00990	.62012
T5 to T2	.02922	.74153	.80858	.71231	.06705	.77936
T7 to T6	.00510	.93825	.95802	.93315	.01978	.95293
T8 to T6	.00728	.93295	.93636	.92567	.00341	.92908

Table 4

Summary of Results of Two-Sample Kolmogorov-Smirnov Tests of Equality of the Cumulative Distribution Functions of Transformed Item Difficulties, for Five Pairs of Aptitude Tests

Test Equating	Maximum Difference in Cumulative Distributions of Equated Deltas	Minimum Difference Needed for Statistical Significance (Alpha=0.001)
T3 to T1	.051	.077
T4 to T1	.066	.077
T5 to T2	.078	.075
T7 to T6	.180	.071
T8 to T6	.239	.071

Table 5

Summary of Results of Two-Sample Kolmogorov-Smirnov Tests of Equality of the Cumulative Distribution Functions of Biserial Correlations for Five Pairs of Aptitude Tests

Test Equating	Maximum Difference in Cumulative Distributions of Biserial Correlations	Minimum Difference Needed for Statistical Significance (Alpha=0.001)
T3 to T1	.096	.077
T4 to T1	.115	.077
T5 to T2	.105	.075
T7 to T6	.123	.071
T8 to T6	.168	.071

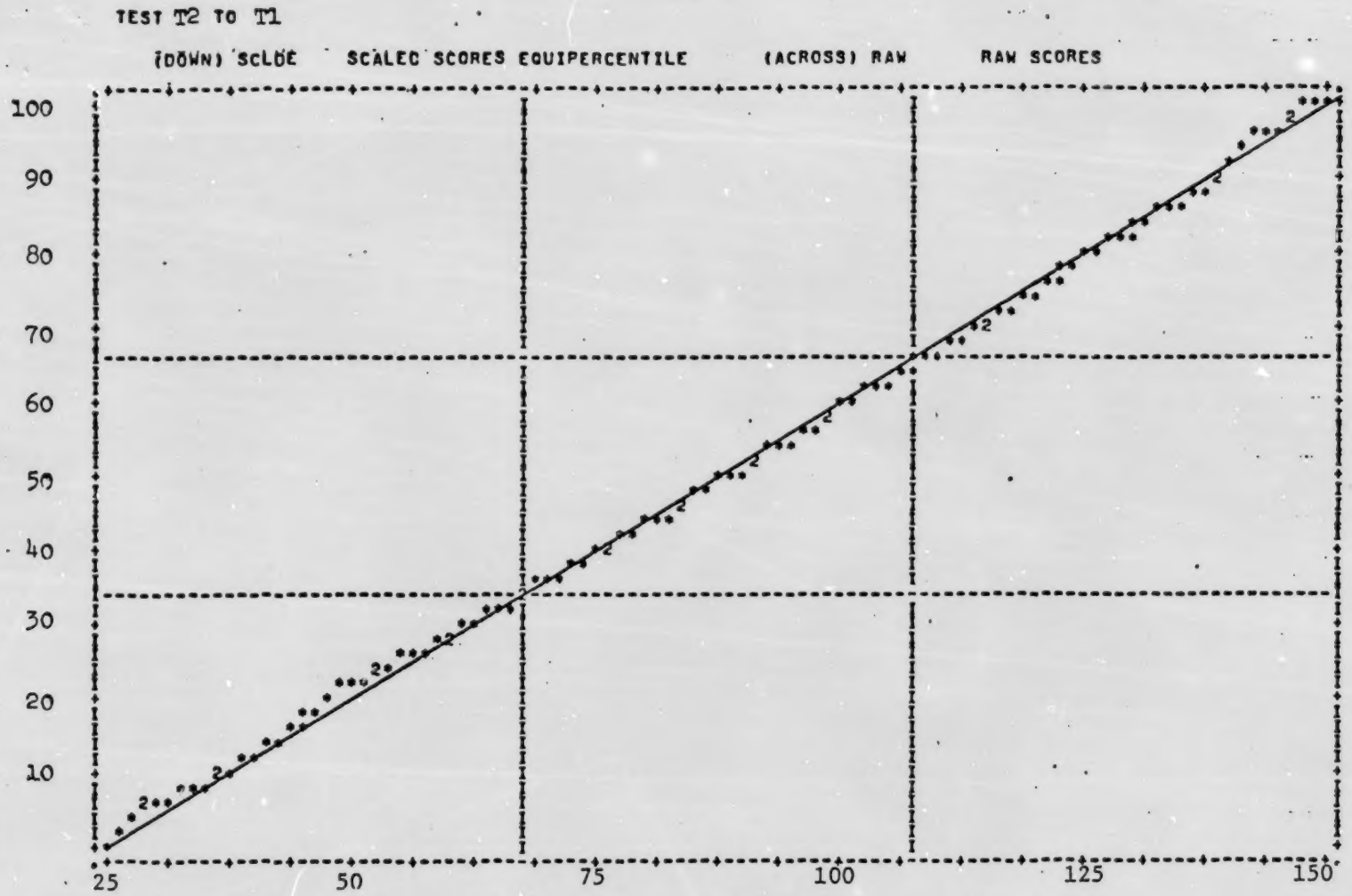

Figure 1. Scatter Diagram of Equipercentile Scaled Scores (vertical) and Raw Scores (horizontal) for the Equating of Test T2 to Test T1.

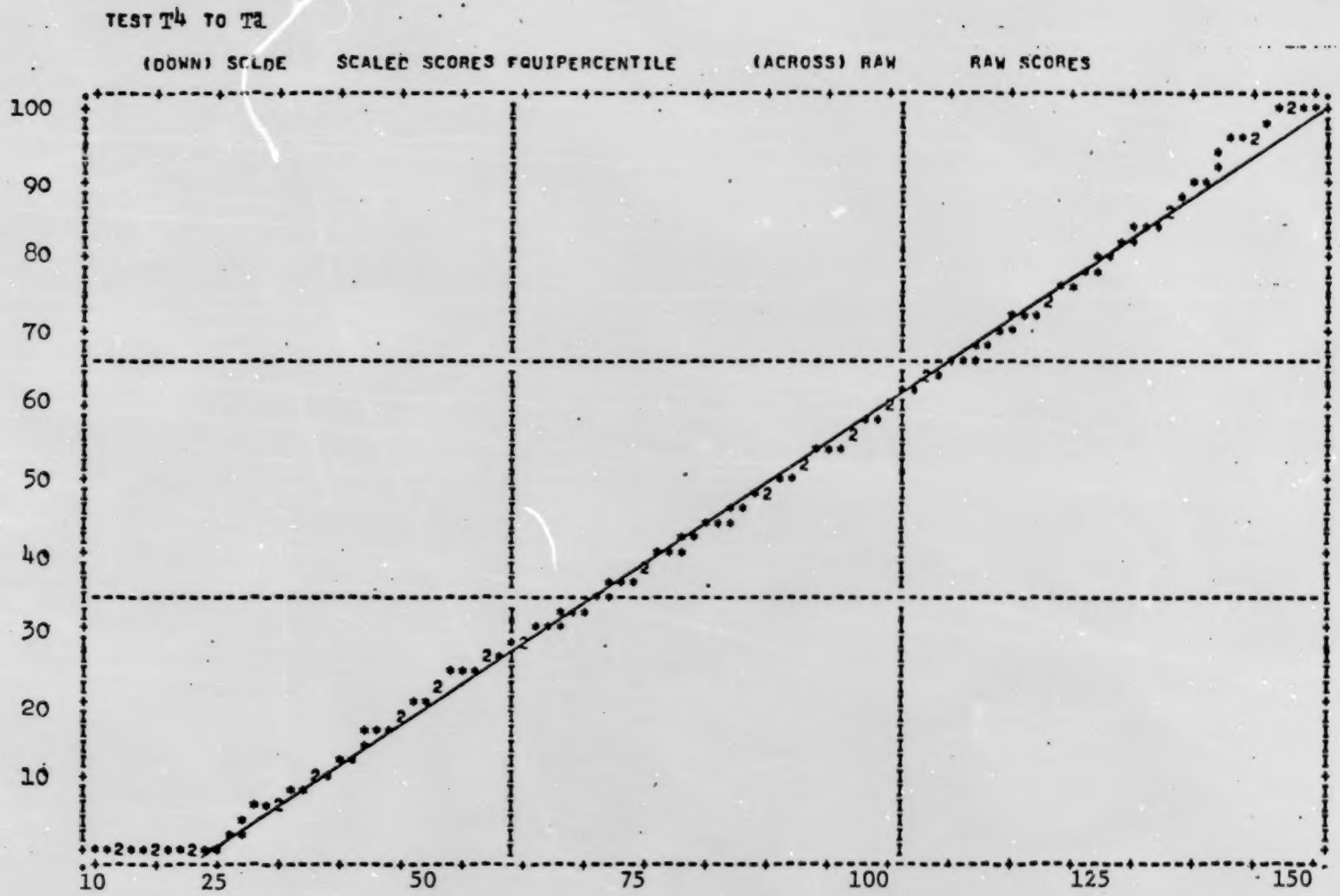

Figure 2. Scatter Diagram of Equipercentile Scaled Scores (vertical) and Raw Scores (horizontal) for the Equating of Test T4 to Test T1 .

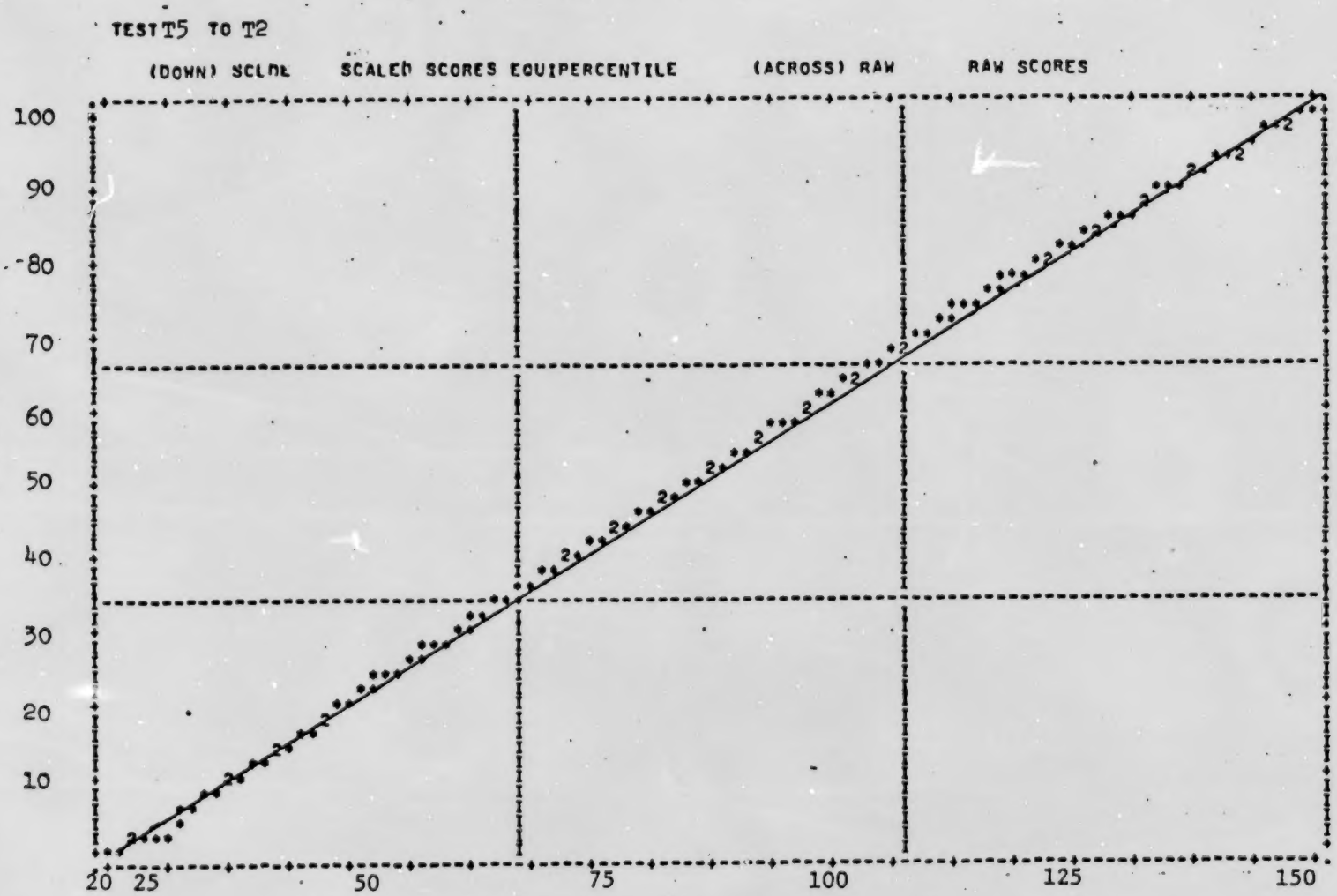

Figure 3. Scatter Diagram of Equipercentile Scaled Scores (vertical) and Raw Scores (horizontal) for the Equating of Test T5 to Test T2.

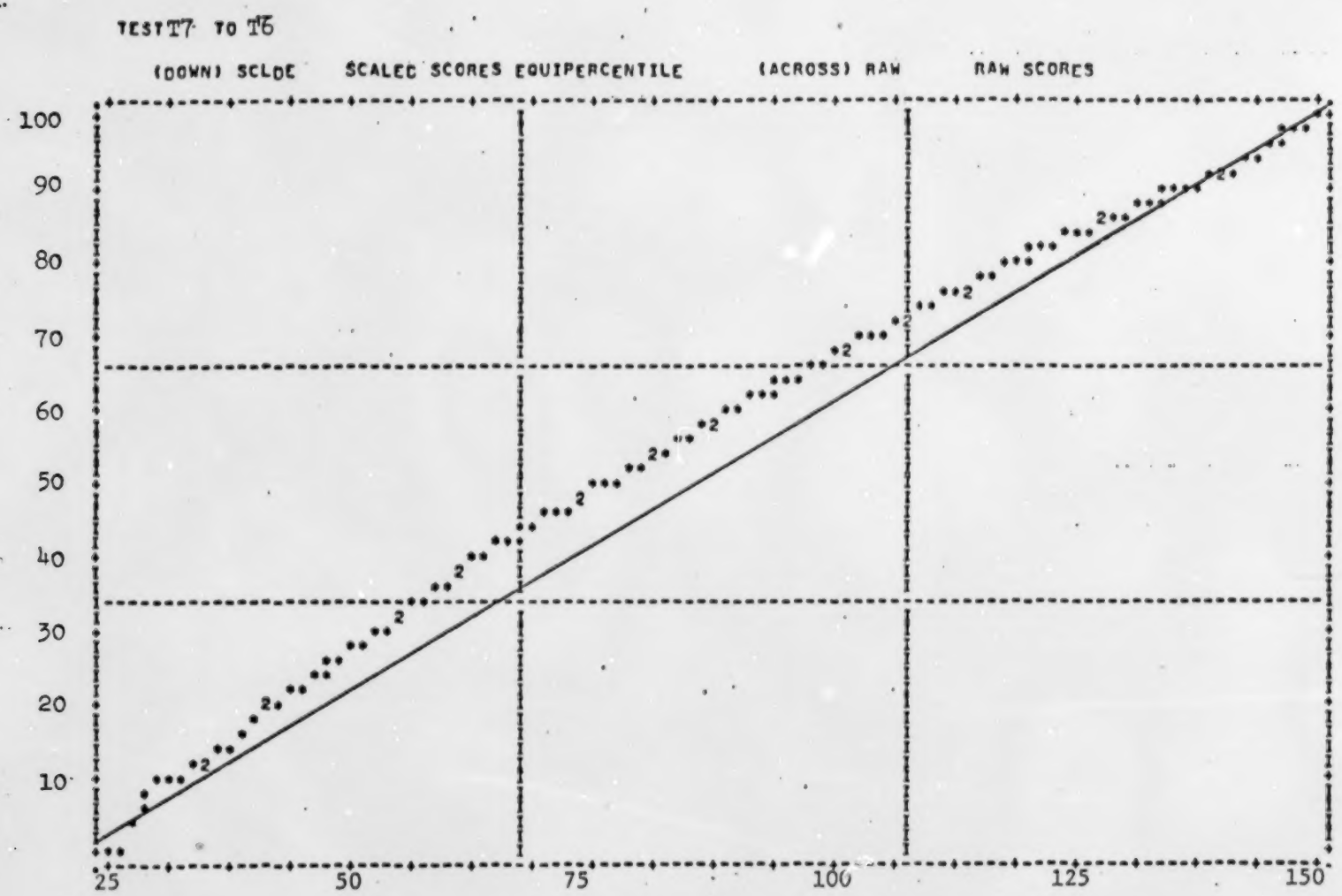

Figure 4. Scatter Diagram of Equipercentile Scaled Scores (vertical) and Raw Scores (horizontal) for the Equating of Test T7 to Test T6.

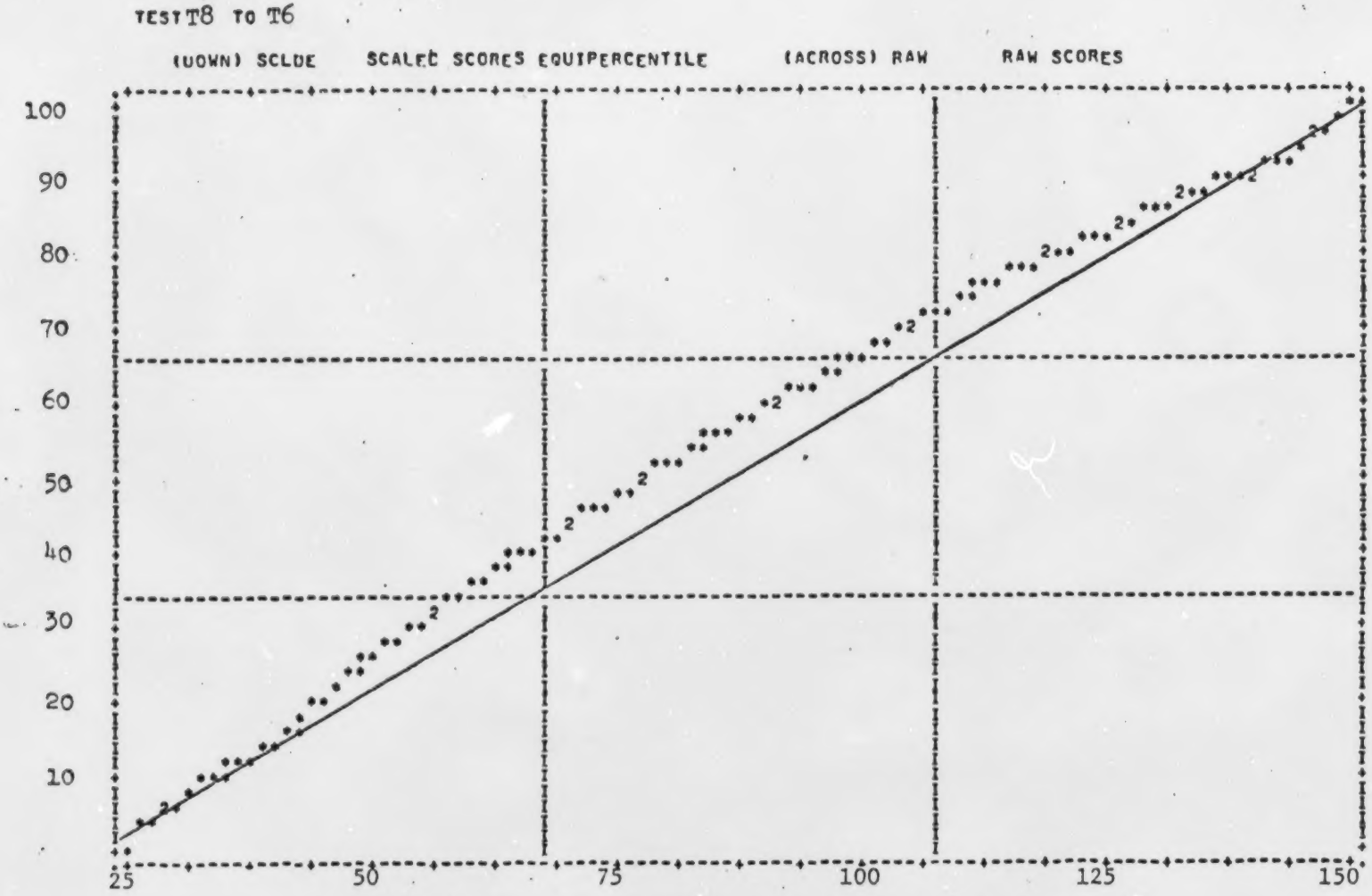

Figure 5. Scatter Diagram of Equipercentile Scaled Scores (vertical) and Raw Scores (horizontal) for the Equating of Test T8 to Test T6.